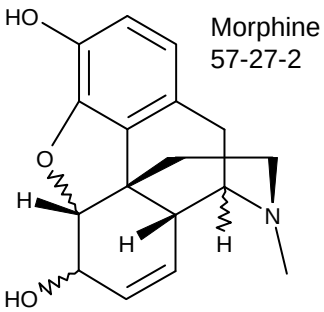
Appendix F. SMILES Notation Tutorial

This is a summary level introduction to SMILES but additional help is available at several online sources including http://www.epa.gov/ncct/dsstox/MoreonSMILES.html#Tutorials.

What is SMILES?
SMILES is the "Simplified Molecular Input Line Entry System," which is used to translate a chemical's three-dimensional structure into a string of symbols that is easily understood by computer software. SMILES notation are used to enter chemical structure into EPI Suite™ estimation programs and ECOSAR. Additional examples of SMILES notations are available in the HELP files of EPI Suite™ and ECOSAR. Software programs are available which can translate a chemical structure into SMILES.

References:
Weininger, D. 1988. SMILES, a Chemical and Information System. 1. Introduction to Methodology and Encoding Rules. J. Chem. Inf. Comput. Sci. 28(1): 31-6.
Wiswesser, W.J. 1954. A Line-Formula Chemical Notation. New York: Cromwell.



Morphine
57-27-2

The purpose of SMILES is to translate the structure to the right, which is Morphine CAS RN 57-27-2, into a linear representation of the molecule so that a computer program can understand the structure.

Here is one SMILES Notation for CAS RN 57-27-2
Oc1ccc2CC(N3C)C4C=CC(O)C5Oc1c2C45CC3

Representing Atoms
Atomic symbols and their corresponding SMILES notations:

C	methane (CH4)	N	ammonia (NH3)
O	water (H2O)	P	phosphine (PH3)
S	hydrogen sulfide (H2S)	Cl	hydrogen chloride (HCl)

Normally hydrogen is not shown.

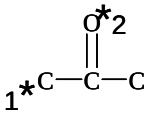
Elements must be shown in brackets: [Au] elemental gold

Representing Bonds
Single, double, triple, and aromatic bonds are represented by the following symbols:
single **–** triple **#** double **=** aromatic **:**
Normally single bonds and aromatic bonds do not need to be written in the SMILES notation.

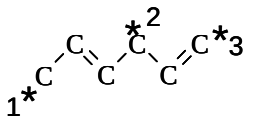
Examples showing bonds are:

CC	ethane (CH3CH3)	C=C	ethylene (CH2=CH2)
COC	dimethyl ether (CH3OCH3)	CCO	ethanol (CH3CH2OH)
C=O	formaldehyde (CH2O)	O=C=O	carbon dioxide (CO2)
O=CO	formic acid (HCOOH)	C#N	hydrogen cyanide (HCN)
[H][H]	molecular hydrogen (H2)		

Bonds in Linear Structures
For linear structures, SMILES notation corresponds to conventional diagrammatic notation except that hydrogen can be omitted. Here are two correct ways to represent Acetone CAS RN 67-64-1, shown here. The numbered asterisks indicate where on the molecule each SMILES string begins. The valid SMILES are: 1. CC(=O)C and 2. O=C(C)C



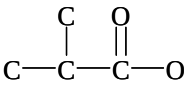
Acetone
67-64-1



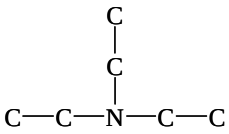
1,4-Hexadiene
592-45-0

Bonds in Linear Structures (continued)
Here are three correct ways to represent 1,4-hexadiene CAS RN 592-45-0. The numbered asterisks indicate where on the molecule each SMILES string begins. The valid SMILES are:
1. CC=CCC=C 2. C(C=C)C=CC 3. CC=CCC=C

Representing Branches
Branches are specified by enclosures in parentheses, for example:



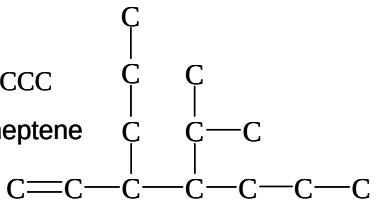
CC(C)C(=O)O
Isobutyric acid
79-31-2



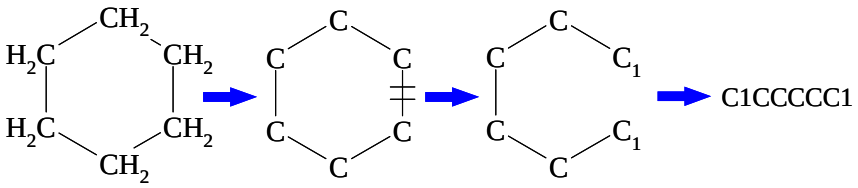
CCN(CC)CC
Triethylamine
121-44-8

Branches also can be nested or stacked, for example:

C=CC(CCC)C(C(C)C)CCC
3-propyl 4-isopropyl 1-heptene

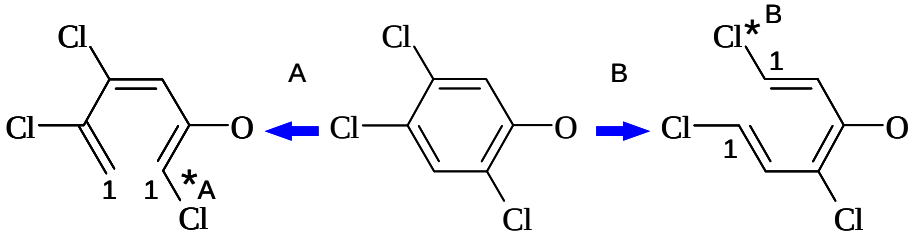


Representing Cyclic Structures
Cyclic structures are represented by breaking one single or double (aromatic) bond in each ring. The bonds are numbered in any order, designating ring-opening/closure bonds by a digit immediately following the atomic symbol at each ring closure. This leaves a connected noncyclic graph, which is written as a noncyclic structure by using the three rules described for atoms, bonds, and branches. A typical example is Cyclohexane CAS RN 110-82-7:



Cyclohexane 110-82-7

Just as in linear structures, there are many different but equally valid descriptions of the same cyclic structure. Many different SMILES notations may be written for the same structure by breaking a ring in different places. For example, two valid SMILES notations for 2,4,5-Trichlorophenol CAS RN 95-95-4 are shown below. The lettered asterisks indicate where on the molecule each SMILES string begins.



SMILES A
c1(Cl)c(O)cc(Cl)c(Cl)c1

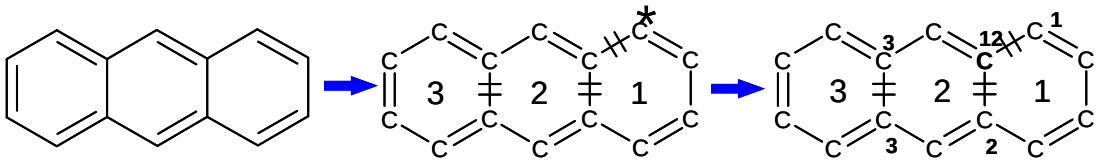
2,4,5-Trichlorophenol
CAS RN 95-95-4

SMILES B
Clc1cc(O)c(Cl)cc1Cl

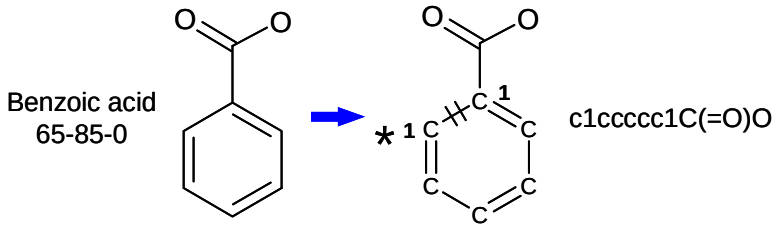
Representing Cyclic Structures (continued)

A single atom may belong to more than one ring and have more than one ring closure. An example of this is Anthracene, in which one atom (bolded below) has more than two ring closures.

Here is the generation of the SMILES notation for Anthracene CAS RN 120-12-7. Number each ring, decide where you want to start the SMILES string (here the SMILES string will begin at the asterisk). Break the rings and give the two atoms at each ring closure the number of that ring.
In this example the SMILES notation for Anthracene is: c1cccc2cc3ccccc3cc12



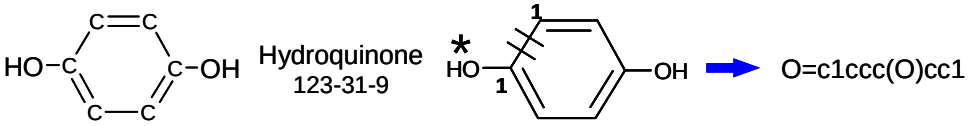
Aromatic structures are distinguished by writing the atoms in the aromatic ring in lower case letters, for example Benzoic acid CAS RN 65-85-0.



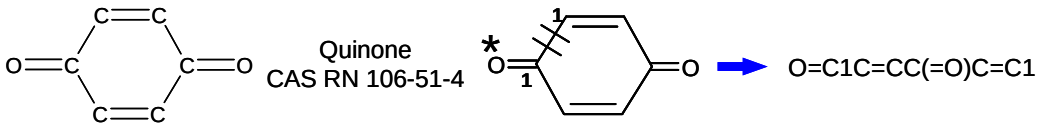
Examples of Aromatic and Nonaromatic Compounds

Of the examples shown on the previous page, Cyclohexane CAS RN 110-82-7 is not aromatic and all carbons are indicated by upper case: C1CCCCC1. Anthracene CAS RN 120-12-7 is aromatic and all carbons are indicated by lower case: c1cccc2cc3ccccc3cc12.

Hydroquinone is **aromatic**. Hydroquinone drawn with aromatic carbons shown in lower case (on the left) and with aromatic carbons hidden (on the right).

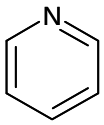


Quinone CAS RN 106-51-4 is **nonaromatic**. Quinone drawn with nonaromatic carbons shown in upper case (on the left) and with nonaromatic carbons hidden (on the right).

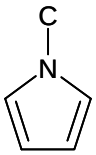


Aromatic Nitrogen

Aromatic nitrogens are specified with the aromatic symbol lower case "n". Examples are pyridine and pyrrole:



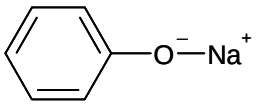
Pyridine
CAS RN 110-86-1
n1ccccc1



Methyl pyrrole
CAS RN 96-54-8
Cn1ccccc1

Disconnected Structures

Disconnected compounds are written as individual structures separated by a period. The order in which ions or ligands are listed is arbitrary. There is no implied paring of one charge with another, and it is not necessary to have a net charge of zero. If desired, the SMILES of one ion may be imbedded in another, as shown in the example the SMILES for Sodium phenoxide.



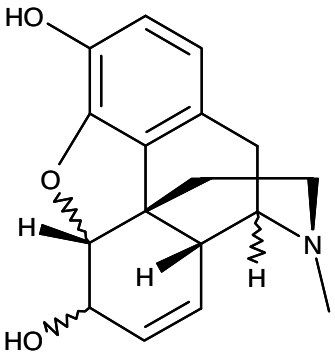
Sodium phenoxide
SMILES Notation

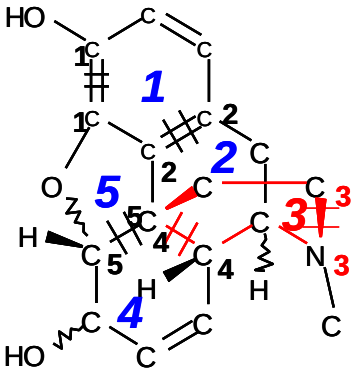
[Na+].[O-]c1ccccc1
or
c1cc([O-].[Na+])ccc1

Evolution of SMILES for Morphine

Here is the generation of one correct SMILES notation for Morphine CAS RN 57-27-2, shown to the right.

Number each ring, decide where you want to start the SMILES string (here the SMILES string will begin at the asterisk). Break the 5 rings and give the two atoms at each ring closure the number of that ring. The dashed line indicates the path followed when this SMILES notation was drawn.





In this example the SMILES notation for Morphine is: Oc1ccc2CC(N3C)C4C=CC(O)C5Oc1c2C45CC3